Copyright © The Author(s) 2027
Jurnal Artificial, Vol 5 No 1, April 2027

Implementation of NLP-Based Hybrid Retrieval Using TF-IDF and Sentence-BERT (SBERT) for Spotlight Search in a Lecturer Certification Video

Evan Januards Sapteno [1], Luther Alexander Latumakulita [2*], Rillya Arundaa [3], Eliasta Ketaren [4], Mahardika Inra Takaendengan [5], Dodisutarma Lapihu [6]

[1, 2, 3, 4, 5, 6] Sam Ratulangi University, Indonesia

* evansapteno16@student.unsrat.ac.id

Abstract

Locating specific information in long educational videos is inefficient when users must inspect the timeline manually. This study implements and evaluates a Natural Language Processing-based hybrid retrieval pipeline for timestamped spotlight search in one Indonesian Lecturer Certification (Sertifikasi Dosen) video. Faster-Whisper medium produced 748 timestamped ASR segments covering approximately 1,788.21 seconds of spoken content. The transcript was converted into 60 retrieval chunks using a fixed 40-second sliding window with a 10-second overlap. Lexical retrieval used TF-IDF, while semantic retrieval used multilingual Sentence-BERT embeddings searched with FAISS IndexFlatIP after L2 normalization. Candidate scores were min-max normalized and combined through α-weighted score fusion. Transcription quality was evaluated using WER and CER, and retrieval was evaluated on 20 queries with 37 relevant query–chunk pairs using Precision@K, Recall@K, MRR@K, and nDCG@K at K = 1, 3, 5, and 10. The ASR evaluation yielded 14.76% WER and 7.23% CER. At α = 0.50, Precision@1 was 0.6000, meaning 12 of 20 queries retrieved a relevant chunk at rank one; Recall@10, MRR@10, and nDCG@10 were 0.9208, 0.7288, and 0.7150, respectively. Among the five tested α settings, α = 0.50 produced the highest Precision@1 and MRR@10, whereas α = 0.75 produced the highest nDCG@10 (0.7273). These findings support the feasibility of the pipeline within the evaluated video and query set, while broader validation is required before generalization.

Keywords: *Automatic Speech Recognition, Hybrid Retrieval, Sentence-BERT, TF-IDF, Timestamped Video Retrieval*

Introduction

Educational videos have become a primary medium for delivering complex material, but their linear structure makes locating specific information less straightforward than searching text documents. When users only need a definition, procedure, assessment component, or example that appears in a particular section, replaying the video from the beginning or manually scanning the timeline is inefficient. Using transcripts allows spoken content to be treated as a corpus that can be indexed and searched. The previous studied showed that educational-video transcripts can be processed into semantic indexes that remain linked to temporal positions (Arazzi et al., 2023). Recent journal studies also show

https://doi.org/10.54065/artificial.5.1.2027.2025

that Whisper-family ASR is being evaluated across non-English and academic speech settings, where transcription quality remains sensitive to language and recording conditions (Polat et al., 2024; Putra et al., 2025). This need is relevant to Lecturer Certification (Serdos) videos because a single video may contain multiple topics, ranging from course descriptions to learning methods, assessment, e-payment, and digital marketing (Novriansyah et al., 2023).

Automatic Speech Recognition (ASR) provides a bridge between audio media and text-based search systems. Whisper was developed as a speech-recognition model trained on large-scale multilingual and multitask data, enabling broad use without dataset-specific fine-tuning (Radford et al., 2023). Recent ASR reviews describe the field's transition toward end-to-end, transformer-based, multilingual, and transfer-learning approaches, while emphasizing that accuracy can vary across languages, speakers, domains, and acoustic conditions (Ahlawat et al., 2025). Journal evaluations of Whisper in non-English settings likewise show that model performance remains empirical rather than uniform across datasets (Polat et al., 2024). For long-form video retrieval, preserving temporal information is also essential because retrieved text must be mapped back to a corresponding video interval. This study therefore uses timestamped ASR output as the textual representation from which retrieval units are constructed.

After transcription, defining the retrieval unit becomes the next issue. If the entire transcript is treated as a single document, location precision is lost because multiple topics are mixed within one representation. Conversely, segments that are too short may separate context that is actually interdependent. Recent journal work on semantic document chunking shows that segmentation strategy affects semantic coherence and downstream retrieval behavior, reinforcing the need to treat chunk size as an experimental design choice rather than a universally fixed value (Kiss et al., 2025). On this basis, this study uses a fixed-size 40-second sliding window with a 10-second overlap. This configuration is not positioned as optimal, but as a fixed setting that enables the pipeline to be evaluated under controlled conditions.

On the retrieval side, lexical and semantic matching provide different signals. TF-IDF assigns greater weight to discriminative terms in a collection and remains a foundational approach for term-weighted retrieval (Salton & Buckley, 1988). Its limitation is the lexical gap: a query and a transcript segment may discuss the same concept using different vocabulary. Surveys of short-text semantic similarity note that sparse lexical evidence, ambiguity, and paraphrasing make semantic matching important for short queries (Han et al., 2021; Amur et al., 2023). Sentence-BERT (SBERT) produces sentence embeddings that can be efficiently compared using vector similarity (Reimers & Gurevych, 2019), while multilingual knowledge distillation extends sentence-level representations across languages (Reimers & Gurevych, 2020). More broadly, neural and pretrained ranking literature describes learned text representations as a major mechanism for semantic retrieval and ranking beyond exact token matching (Guo et al., 2020; Fan et al., 2022; Wang et al., 2024); Riskawati & Saleh, 2026.

Nevertheless, semantic retrieval cannot be assumed to always replace lexical retrieval. The BEIR benchmark demonstrates substantial variation across domains and motivates heterogeneous evaluation of retrievers (Thakur et al., 2021). Journal-based reproducibility analyses similarly report that lexical and neural retrieval models can exhibit different strengths across collections, languages, and topics, so gains should be interpreted within

the tested corpus and evaluation protocol (Marchesin et al., 2020). The further Studied showed that hybrid score fusion must be interpreted together with score normalization and parameter tuning (Bruch et al., 2024). These findings support the use of hybrid retrieval in this study while also requiring claims to remain restricted to the experimental conditions actually evaluated.

The need for timestamped search has also been studied from the perspective of video retrieval. Survey work on video moment localization defines the task as locating a target temporal segment from a natural-language query and reviews the importance of aligning language with temporal spans (Liu et al., 2023). More recent work on video temporal grounding continues to emphasize fine-grained temporal localization and evaluation protocols for natural-language video search (Wu et al., 2026). The previous studied provided an instructional-video dataset that associates questions with relevant video segments (Gupta et al., 2023). While the another studied showed that timestamped ASR transcripts can connect text search with directly playable video clips (Masuyama et al., 2024). These directions emphasize that language representations and temporal information can be combined to support video navigation. However, the focus of this study is narrower: the search units are timestamped transcript chunks extracted from one Indonesian-language Serdos video.

In transcript-based video retrieval, the quality of navigation depends on the interaction among several stages rather than on one retrieval model in isolation. ASR determines which lexical and semantic evidence becomes available to the search system, timestamp preservation determines whether a retrieved passage can be mapped back to the correct video interval, and chunking determines the amount of context represented by each retrieval unit. Retrieval evaluation therefore needs to distinguish between the ability to place one useful segment at the top of the ranking and the ability to cover the wider set of relevant segments. This distinction is especially important for educational videos, where a single question may be answered by one concise segment or by several neighboring segments that jointly contain the required explanation. Prior work on transcript indexing and timestamped video access shows the value of linking text retrieval to temporal navigation (Arazzi et al., 2023; Masuyama et al., 2024). While modern IR evaluation emphasizes interpreting ranking metrics according to the behavior they measure rather than collapsing performance into a single score (Moffat, 2022).

Based on the reviewed studies, research that integrally evaluates a pipeline comprising transcription, timestamped chunking, lexical retrieval, semantic retrieval, vector search, score fusion, and ranking evaluation for segment search in Indonesian-language Serdos educational videos remains limited. This study also incorporates semi-automatically constructed ground truth based on transcript-content auditing, α-parameter sensitivity analysis, a comparison of TF-IDF, SBERT, and hybrid weighting configurations, and error analysis of failed queries. The contribution is not the creation of new TF-IDF, SBERT, FAISS, or weighted score fusion algorithms; rather, it lies in implementing and evaluating this pipeline in a specific data context. Reproducibility research in IR emphasizes traceable experimental configurations, cross-collection validation, and care when interpreting aggregate effectiveness scores (Marchesin et al., 2020; Maistro et al., 2023).

Accordingly, this study aims to implement a spotlight retrieval system that accepts user queries and returns the most relevant transcript chunks together with their timestamps; evaluate transcription quality using WER and CER; and measure retrieval

performance using Precision@K, Recall@K, MRR@K, and nDCG@K at K = 1, 3, 5, and 10. The study also evaluates five α values in weighted score fusion, compares TF-IDF, SBERT, and hybrid configurations within the same candidate framework, and identifies error patterns at Top-1 and Top-3. The main research question is the extent to which the integration of lexical and semantic retrieval can support timestamped segment search in the evaluated video, and what limitations are revealed by the ranking results and error analysis.

Method

Research Design and Data

This study uses an experimental implementation-and-evaluation design on one publicly available Indonesian-language Lecturer Certification educational video from YouTube. The analysis focuses on audio content transcribed into timestamped text; visual content such as slides, diagrams, images, and unspoken on-screen text is not included in the retrieval representation. The entire implementation was conducted in Google Colab. A single video was selected so that acquisition, transcription, segmentation, indexing, retrieval, and evaluation could be traced within a controlled corpus; consequently, the findings are treated as results for this scenario rather than as general performance estimates for all educational videos.

The video audio was transcribed using Faster-Whisper medium. The ASR output contains start time, end time, and text. ASR evaluation covers the entire spoken content, approximately 1,788.21 seconds and 748 segments. The reference transcript was prepared by listening again to the spoken content and correcting differences in the automatic transcription. Thus, WER and CER were calculated over the complete spoken content used as the basis for index construction.

Table 1. *Summary of research data and pipeline configuration*

Component	Configuration/Value
Data source	1 Lecturer Certification educational video from YouTube
Language	Indonesian
ASR model	Faster-Whisper medium
Spoken-content coverage	±1,788.21 seconds; 748 ASR segments
Chunking	Fixed-size sliding window 40 seconds, 10-second overlap, 30-second stride
Number of retrieval units	60 timestamped chunks
Lexical retrieval	TF-IDF; index matrix 60 × 3,394
Semantic retrieval	Sentence-BERT paraphrase-multilingual-mpnet-base-v2; embeddings 60 × 768
Vector search	FAISS IndexFlatIP after L2 normalization
Score fusion	Min-max normalization + weighted score fusion
Retrieval evaluation	20 queries; 37 relevant query–chunk pairs; K = 1, 3, 5, 10

Table 1 summarizes the configuration used consistently throughout the experimental process. The 60 chunks resulted from segmenting the complete transcript using a 40-second window with a 10-second overlap. TF-IDF and SBERT representations were built over the same set of chunks so that score differences reflect the characteristics of the lexical and semantic paths rather than differences in document units.

Transcript Chunking and Index Construction

Raw ASR segments have non-uniform durations and do not necessarily contain sufficient context to answer a query. Therefore, the transcript was formed into time-based chunks. The 40-second window was shifted every 30 seconds so that the final 10 seconds of one chunk reappeared in the next chunk. The overlap was intended to reduce the risk of information being split exactly at a segmentation boundary, but its consequence is increased content similarity between adjacent chunks. The 40/10 configuration was applied consistently and was not compared with other sizes; therefore, this study does not claim that the configuration is optimal.

On the lexical path, TfidfVectorizer was used to construct TF-IDF representations with the basic normalization provided by the vectorizer pipeline. On the semantic path, chunk text was processed directly using the paraphrase-multilingual-mpnet-base-v2 model. This model was selected because the study uses Indonesian queries and transcripts and requires multilingual sentence-embedding representations. Chunk and query embeddings were L2-normalized before vector search. FAISS IndexFlatIP was used for exact inner-product search; for unit vectors resulting from L2 normalization, inner product is equivalent to cosine similarity. The role of FAISS is limited to vector indexing and search, while the semantic representation is produced by SBERT (Douze et al., 2026).

Hybrid Score Fusion

TF-IDF and SBERT produce different score ranges. To combine the two signals on a more comparable scale, candidate scores from each path were normalized separately for each query using min-max normalization. If all candidates on one path have identical scores, the implementation assigns a normalized value of 0 to avoid division by zero. The normalized values are relative to the candidate set for the same query and are not interpreted as probabilities of relevance.

After normalization, the scores were combined using weighted score fusion. The α parameter controls the contribution of SBERT, while $(1 - \alpha)$ controls the contribution of TF-IDF. An α value of 0 means that weighting comes only from TF-IDF, $\alpha = 1$ means that it comes only from SBERT, and intermediate values combine the two signals. The use of weighted fusion follows the principle that lexical and semantic signals can provide complementary evidence, while the effectiveness of the combination still depends on score calibration and the target collection (Bruch et al., 2024; Marchesin et al., 2020).

$$Score_norm(c) = (Score(c) - Score_min) / (Score_max - Score_min)$$

$$FinalScore(c) = (1 - \alpha) \times TFIDF_norm(c) + \alpha \times SBERT_norm(c)$$

These two equations represent the main operations in the fusion stage. Normalization is performed before weighting so that changes in α primarily modify the relative contribution of the two paths to the final ranking rather than merely following the path with the larger raw-score scale.

System Implementation and Retrieval Procedure

The implementation was developed in Python on Google Colab and organized so that index construction and query-time retrieval could be distinguished operationally. During index construction, the video audio was processed by Faster-Whisper, the timestamped ASR segments were aggregated into 40-second chunks with a 10-second overlap, and each

chunk retained its identifier, start time, end time, and transcript text. The same 60 chunks were then represented in two forms. The lexical path stored TF-IDF vectors, while the semantic path encoded each chunk with paraphrase-multilingual-mpnet-base-v2, applied L2 normalization, and stored the resulting embeddings in FAISS IndexFlatIP. These reusable artifacts allowed subsequent queries to be processed without repeating transcription or chunk generation.

At query time, the text query was represented through both retrieval paths. TF-IDF produced lexical similarity scores against the chunk corpus, while SBERT produced a normalized query embedding that was searched against the FAISS index. Candidate identifiers returned by the two paths were combined into a unique candidate set. For those candidates, lexical and semantic scores were retained and independently min-max normalized before weighted score fusion. The fused score was then used to sort the candidate set, after which the requested Top-K chunks were returned with their timestamps and transcript excerpts. This procedure preserves a clear separation between representation, candidate retrieval, score calibration, and final ranking. It also explains why the baseline comparison in this study should be interpreted as a comparison of weighting configurations within a shared candidate framework rather than as three completely independent retrieval systems.

The research and evaluation workflow was kept separate from the Gradio demonstration notebook. The research notebook handled transcription, chunking, index construction, ground-truth preparation, ASR evaluation, retrieval evaluation, baseline comparison, α testing, and error analysis. The demonstration notebook loaded the resulting retrieval artifacts and provided the interactive query interface. Consequently, the interface did not redefine the evaluation protocol or generate the reported metrics. Figure 1 summarizes the experimental retrieval pipeline used to transform the video into timestamped Top-K results.

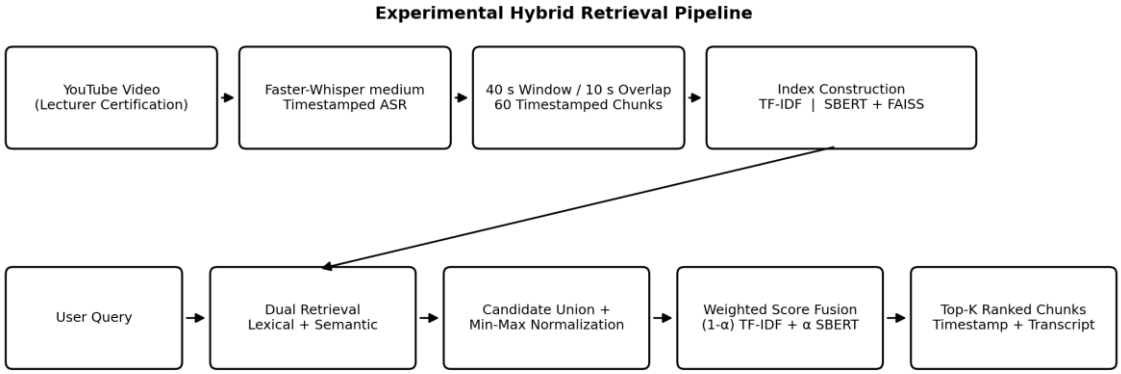


Figure 1. *Experimental hybrid retrieval pipeline for timestamped video segment search*

As shown in Figure 1, processing proceeds from the source video to timestamped ASR output and fixed-size chunks before the corpus is indexed through parallel lexical and semantic representations. User queries then enter the same two retrieval paths, after which the candidate sets are merged, normalized, and fused into a final ranking. The timestamp stored with each chunk is preserved through the ranking stage so that a retrieved text segment remains directly traceable to the corresponding interval in the original video.

Ground Truth Construction

The ground truth was constructed semi-automatically through transcript-content auditing. Twenty queries were designed to cover information from the beginning, middle, and end of the video, including lecturer identity, course context, learning methods, schedule, assessment, e-payment, digital marketing, social media, and assignments. For each query, a reference time span was used to form the initial chunk candidates. The system then added one neighboring chunk on the left and right to account for information crossing sliding-window boundaries.

Final relevance decisions were not determined solely by temporal proximity. Each candidate was examined against the transcript content and labeled relevant if the chunk answered or supported the information need expressed by the query. Relevance was binary; a chunk that discussed a similar topic but did not answer the query was labeled non-relevant. This procedure produced 37 relevant query–chunk pairs, or an average of 1.85 relevant chunks per query. The final audit was conducted by one researcher and did not involve a second independent annotator, so potential subjectivity in relevance judgments remains a limitation of the study.

Evaluation Metrics and Experimental Scenarios

ASR quality was evaluated using WER and CER. WER counts substitutions, deletions, and insertions at the word level against the reference transcript, whereas CER applies the same edit-distance principle at the character level. Recent reviews of ASR evaluation confirm the continued centrality of word-error measures and the complementary use of character-level error measures in multilingual settings (Ahlawat et al., 2025; Alabbad & Alhoshan, 2026). Retrieval was evaluated using macro-averaged Precision@K, Recall@K, MRR@K, and nDCG@K at K = 1, 3, 5, and 10. Each query has equal weight in the macro average. Precision@1 has a specific interpretation under binary relevance: for one query, the value is 1 if the first result is relevant and 0 otherwise, so macro Precision@1 equals the proportion of queries that successfully place a relevant chunk at rank one.

Recall@K measures the proportion of all relevant chunks found up to cut-off K. MRR@K evaluates the position of the first relevant chunk through reciprocal rank, whereas nDCG@K evaluates ranking quality with a larger discount for lower positions (Järvelin & Kekäläinen, 2002; Moffat, 2022). Because these three metrics serve different functions, performance is not interpreted from a single value alone.

α sensitivity was tested at 0.00; 0.25; 0.50; 0.75; and 1.00 using the same queries and ground truth. Precision@1, Recall@10, MRR@10, and nDCG@10 were used to compare ranking changes. The baseline comparison used three configurations within the same candidate framework: α 0.00 for TF-IDF weighting, α 1.00 for SBERT weighting, and α 0.50 as the main hybrid configuration. In this study, the term baseline refers to weighting configurations within the same candidate framework rather than experiments with fully independent candidate-generation paths.

Error analysis was conducted on queries that failed at Top-1 and false positives at Top-3. Its purpose was to distinguish failures to place a relevant result first from failures to retrieve all relevant chunks, and to identify patterns such as topical similarity, distributed information, compound queries, or rankings that assign high scores to chunks that remain related to the topic but do not answer the information need.

Results

ASR and Chunking Results

The ASR process produced 748 timestamped segments covering approximately 1,788.21 seconds of spoken content. All segments were then reorganized into 60 chunks using a fixed-size 40-second sliding window with a 10-second overlap. The resulting TF-IDF index had dimensions of 60 × 3,394, while the SBERT embeddings had dimensions of 60 × 768 and the FAISS index contained 60 vectors.

Table 2. *ASR quality evaluation results*

Component	Result
ASR model	Faster-Whisper medium
Evaluation segments	748 segments
Hypothesis words	3,455
Reference words	3,678
WER	0.1476 (14.76%)
CER	0.0723 (7.23%)

Table 2 shows that the automatic transcription is not identical to the reference transcript. WER of 14.76% and CER of 7.23% are used as context for the quality of the retrieval input data; these values are not converted into a claim that the transcript is error-free. All TF-IDF and SBERT representations in the experiment were built from this ASR transcript.

Retrieval Performance at Different K

Retrieval evaluation used 20 queries and 37 relevant query–chunk pairs. Under the main α 0.50 configuration, the metric values changed with ranking depth as shown in table.

Table 3. *Retrieval evaluation results by K value*

Metric	K=1	K=3	K=5	K=10
Precision@K	0.6000	0.3833	0.2800	0.1600
Recall@K	0.3458	0.6625	0.8208	0.9208
MRR@K	0.6000	0.7000	0.7225	0.7288
nDCG@K	0.6000	0.6134	0.6782	0.7150

At K=1, Precision@1 of 0.6000 means that 12 of 20 queries placed a chunk included in the relevant ground truth at rank one, while eight queries failed at Top-1. At K=3, 17 of 20 queries had at least one relevant chunk among the top three results. Recall increased from 0.3458 at K=1 to 0.9208 at K=10. MRR increased most substantially from K=1 to K=3, from 0.6000 to 0.7000, and then increased only to 0.7288 at K=10. nDCG increased from 0.6000 to 0.7150 when the cut-off was extended to ten results.

α Sensitivity

α testing changes the contribution weights of TF-IDF and SBERT without changing the set of 20 queries, the ground truth, or the candidate mechanism. The results for the five configurations are shown in Table 4.

Table 4. *α-parameter sensitivity across the five tested configurations*

α	Characteristics	Precision@1	Recall@10	MRR@10	nDCG@10
0.00	100% TF-IDF	0.4500	0.8750	0.6026	0.6191
0.25	Hybrid leaning toward TF-IDF	0.5500	0.8958	0.6700	0.6663
0.50	Balanced hybrid	0.6000	0.9208	0.7288	0.7150
0.75	Hybrid leaning toward SBERT	0.5500	0.9208	0.7264	0.7273
1.00	100% SBERT	0.4500	0.9208	0.6496	0.6790

Among the five tested values, α 0.50 produced the highest Precision@1 (0.6000) and the highest MRR@10 (0.7288). Recall@10 of 0.9208 was obtained at α 0.50, 0.75, and 1.00. The highest nDCG@10 was instead obtained at α 0.75, with a value of 0.7273. Therefore, no single α produced the highest value on every metric.

Baseline Comparison

The baseline comparison was conducted within the same candidate framework and ground truth, using the extreme α values as single-weight configurations and α 0.50 as the hybrid configuration. The MRR@10 and nDCG@10 results are shown in Table 5.

Table 5. *Comparison of baseline retrieval configurations*

Configuration	MRR@10	nDCG@10
TF-IDF (α=0.00)	0.6026	0.6191
SBERT (α=1.00)	0.6496	0.6790
Hybrid (α=0.50)	0.7288	0.7150

In this scenario, Hybrid α 0.50 had an MRR@10 that was 0.0792 higher than SBERT and 0.1262 higher than TF-IDF. For nDCG@10, the differences were 0.0360 relative to SBERT and 0.0959 relative to TF-IDF. This comparison describes only the three weighting configurations under the same video, queries, ground truth, and candidate mechanism.

Representative Error Cases

Precision@1 of 0.6000 leaves eight queries that failed to place a relevant chunk at rank one. Several representative examples are shown in Table 6. This table does not change the ground-truth labels; the observed-pattern column summarizes the qualitative audit reported in the final thesis.

Table 6. *Examples from the error analysis of queries that failed at Top-1*

Query (shortened)	Top-1	Relevant chunk	Observed pattern
Lecturer identity and university	1	0	Adjacent introductory context; Top-1 is not the segment containing the complete answer
E-commerce learning method	3	0, 1	A same-topic adjacent chunk received a higher score
Main course objective	2	3	Relevant information is in a neighboring chunk; high topical similarity
Types of digital marketing	22	23, 24, 25, 26	A general digital-marketing definition/topic outranked the requested list of types
Communication with Gen Z	38	51	General marketing context is more semantically similar than the specific answer
Assignment submission medium and next activity	10	57	Schedule/learning terms occur elsewhere but do not answer the query

The error analysis shows that a Top-1 failure is not always equivalent to failing to find a relevant chunk at greater ranking depths. In several cases, the relevant chunk still appeared at a subsequent rank. False positives primarily occurred when chunks contained vocabulary or semantic context that remained related to the query but did not directly satisfy the main information need.

Discussion

ASR Quality as Retrieval Input

WER of 14.76% and CER of 7.23% indicate that the input transcript still contains substitutions, deletions, and insertions, although the character-level error rate is lower than the word-level error rate. In the retrieval pipeline, the most important consequence is not merely the magnitude of WER, but whether the errors affect terms that serve as query signals. TF-IDF depends directly on token occurrence; mistranscription of terms such as method names, platforms, or concepts can reduce lexical matching. SBERT may be more tolerant of surface variation when sentence context still preserves meaning, but embeddings do not guarantee correct information recovery when important keywords are missing or altered. Therefore, retrieval results should be interpreted as the performance of a system using a transcript with this empirically measured quality, not as though it were using ground-truth text.

These results are consistent with the role of ASR as an upstream componentthe previous studied demonstrated the broad multilingual capability of Whisper (Radford et al. 2023). while recent ASR surveys stress that performance still varies with language, speaker, domain, and recording conditions (Ahlawat et al., 2025). Journal evaluations of Whisper in Turkish and Indonesian academic audio likewise report dataset-dependent WER values, reinforcing the need to measure transcription quality in the target setting rather than assume uniform accuracy (Polat et al., 2024; Putra et al., 2025). This study extends that use not by proposing a new ASR model, but by treating timestamped transcripts as a corpus for retrieval and retaining WER/CER as indicators of input-data quality.

Precision–Recall Trade-off Across K

The clearest pattern in Table 3 is the decrease in Precision@K alongside the increase in Recall@K as K becomes larger. This is not a contradiction. At K=1, the system has only one position that can be correct or incorrect; Precision@1 of 0.6000 means that 60% of queries obtain a hit at the first position. When the list is expanded to ten results, the system has more opportunities to include relevant chunks, so Recall@10 rises to 0.9208. At the same time, each query has only 1.85 relevant chunks on average in the ground truth. In aggregate, there are 37 relevant pairs for 20 queries, so even if every relevant chunk were retrieved in Top-10, the maximum relevance density across 200 result slots would be 37/200 = 0.1850. In this context, Precision@10 of 0.1600 should not be read as a statement that the system is only "16% accurate," but rather as a consequence of a cut-off that is far larger than the number of relevant chunks per query.

The combination of P@1 and R@10 also provides a more useful operational view. P@1 evaluates whether users immediately receive the correct segment, whereas R@10 evaluates whether the broader result list covers most of the relevant evidence. For a spotlight function, Top-1 placement remains important because users expect rapid access. However, high recall at Top-10 indicates that many first-rank failures could be addressed by improving the ordering rather than always expanding the candidate corpus. This perspective is consistent with the previous studied, who emphasizes that IR metrics should be interpreted according to what they measure rather than reduced to a single performance label (Moffat, 2022).

MRR and nDCG Capture Different Ranking Behaviors

MRR increased from 0.6000 at K=1 to 0.7000 at K=3, and then only to 0.7288 at K=10. The diminishing increase indicates that extending the list from five to ten results changes the position of the first relevant chunk only slightly on average. By contrast, nDCG increased more noticeably from 0.6782 at K=5 to 0.7150 at K=10 because this metric considers the distribution of all relevant chunks in the ranking and applies a position-based discount. Under binary relevance, nDCG remains sensitive to whether additional relevant chunks appear earlier or later in the list (Järvelin & Kekäläinen, 2002).

This difference in function explains why α 0.50 can have a slightly higher MRR@10 than α 0.75, while α 0.75 has the higher nDCG@10. Among the five tested configurations, α 0.50 performs better for the objective of "how quickly the first relevant hit appears," whereas α 0.75 produces a slightly better ordering of all relevant hits according to nDCG. Parameter selection should therefore depend on the application objective. A system that strongly prioritizes the first response may emphasize MRR/P@1, whereas a system that displays several results at once may place greater emphasis on nDCG and Recall.

Sensitivity to α Weighting

α testing shows that small changes in the lexical and semantic contributions can alter the ranking. P@1 rose from 0.4500 at α 0.00 to 0.6000 at α 0.50, then fell again to 0.4500 at α 1.00. MRR@10 showed a similar pattern and reached 0.7288 at α 0.50. However, nDCG@10 reached its highest value at α 0.75 rather than α 0.50. The absence of a single configuration that is highest across all metrics is the main reason α 0.50 is not described as optimal.

Conceptually, these results are consistent with the hybrid-retrieval literature. The previous studied showed that fusion effectiveness is affected by normalization and weighting choices (Bruch et al., 2024). While reviews of neural and pretrained IR emphasize that semantic ranking models and traditional lexical signals retain different strengths depending on the retrieval setting (Guo et al., 2020; Fan et al., 2022; Wang et al., 2024). In this study, α = 0.50 was selected as the main configuration because it had the highest P@1 and MRR@10 among the five tested values, not because there is evidence that a 50:50 weighting will outperform alternatives on other videos or queries.

There is an additional methodological limitation: the same 20 queries were used both to compare α values and to report final performance. Thus, configuration selection and evaluation are not fully independent. The selected configuration may therefore be overly adapted to the characteristics of the available queries. IR reproducibility studies emphasize the importance of traceable configurations, independent replication, and evaluation procedures that do not reduce reproducibility to one aggregate score alone (Marchesin et al., 2020; Maistro et al., 2023). Future studies should use a validation set to select α and a separate test set to estimate final performance.

Hybrid and Single-Weight Baseline Comparison

Within this experimental framework, hybrid α 0.50 achieved MRR@10 of 0.7288 and nDCG@10 of 0.7150, higher than the compared TF-IDF and SBERT weighting configurations. This difference supports the idea that two retrieval signals can assist each other for certain queries. Lexical matching is useful when a query contains terms that also appear in the transcript, whereas SBERT can help when semantic relationships do not

depend on identical token matches. However, this baseline must be interpreted precisely: TF-IDF and SBERT are represented by α 0 and 1 within the same candidate-generation mechanism. Because candidates were not built through two fully independent experiments, the results constitute a comparison of weighting configurations within the same pipeline.

These findings also do not justify a claim that hybrid retrieval is always superior. BEIR shows that retriever performance can vary substantially across domains (Thakur et al., 2021). While journal surveys of BERT-based and neural IR show that semantic retrieval effectiveness depends on task formulation, document characteristics, training assumptions, and the balance between exact and contextual signals (Guo et al., 2020; Wang et al., 2024). In this study, the error analysis specifically reveals cases in which semantic and lexical similarity are both high for chunks that are still incorrect with respect to the information need. Therefore, the numerical advantage of the hybrid configuration is limited to one video, 20 queries, 37 relevant pairs, and the tested configurations.

Error Patterns and the Role of Chunking

The eight Top-1 failures show that the main challenge is not merely finding the topic, but distinguishing chunks that actually answer the query from chunks that discuss a similar context. For example, a query about types of digital marketing ranked a chunk discussing the definition or general context of digital marketing highly, even though the ground truth required the segment containing the list of types. Similar patterns appeared for queries about course objectives, learning methods, and communication with Gen Z. In such cases, embeddings may capture broad semantic proximity, while TF-IDF may reinforce shared terms, but neither has an explicit mechanism for checking whether all components of the question have been answered.

Fixed-size chunking contributes to both sides of the issue. The 10-second overlap helps preserve context around boundaries so that an answer crossing a segmentation point is not immediately lost. On the other hand, overlap causes neighboring chunks to share some text and increases the likelihood that two chunks receive nearly identical scores. Recent journal research on semantic chunking likewise shows that segmentation choices materially affect the coherence of retrieval units and downstream retrieval quality (Kiss et al., 2025). Because this study tests only the 40/10 configuration, it cannot determine whether changing chunk size or overlap would improve P@1. Multi-configuration experiments are needed to isolate the effect of segmentation on ranking errors.

Ranking Error Interpretation and Relevance Granularity

The qualitative errors further clarify the distinction between topical similarity and answer relevance. In the thesis-level audit, every query had at least one Top-3 result that was not included in the binary ground truth, even though relevant results were also present for many queries. These false positives were not random: they commonly occurred when a chunk shared important vocabulary with the query, discussed the same broad topic, or was temporally adjacent to the answer, but did not contain the specific information requested. A query asking for assessment components, for example, can assign high lexical or semantic similarity to a chunk that discusses the course generally without stating the assessment items and weights. This pattern shows that similarity-based retrieval can identify topical neighborhoods without necessarily verifying that all elements of the information need have been satisfied. The same limitation becomes more visible for

compound questions, because a chunk can match one clause strongly while omitting another. The 10-second overlap also creates near-neighbor chunks with partially repeated content, increasing the chance that two semantically similar candidates exchange positions near the top of the ranking. These observations support future evaluation of reranking methods that explicitly model query–chunk relevance after candidate generation, while also motivating experiments with multiple chunk sizes and overlaps. Such extensions should be tested on separate validation and test queries so that improvements in Top-1 ranking are not conflated with parameter adjustment on the final evaluation set.

Implications, Reproducibility, and Limitations

In practical terms, the findings show that timestamped transcripts can serve as a navigation layer over educational videos. The system can return time spans and text excerpts so that users do not have to begin searching from the first minute. This approach is consistent with ASR-based text-to-video navigation (Arazzi et al., 2023; Masuyama et al., 2024) and with journal surveys of natural-language video moment localization and temporal grounding (Liu et al., 2023; Wu et al., 2026). For learning scenarios, the primary benefit is not to replace the video, but to shorten the path to segments that may answer the user's information need.

Nevertheless, the experimental scope remains narrow. First, only one Serdos video was used, so variation in speakers, audio quality, topics, duration, and presentation style is not represented. Second, only 20 queries with 37 relevant pairs were evaluated. Third, the ground truth was validated by one researcher without a second independent annotator and used binary relevance; therefore, variation in human judgments and degrees of relevance were not measured. Fourth, the 40-second chunks with a 10-second overlap were not compared with other configurations. Fifth, the same queries were used for α selection and final reporting. Sixth, visual content in the video was not analyzed. Seventh, Gradio functions only as a demonstration prototype in a separate notebook and is not a production application with a backend, database, or permanent hosting.

These limitations restrict generalizability. The results should be interpreted as evidence of implementation feasibility within the evaluated experimental framework. Evaluation on more videos, more varied queries, independent annotators, and a separate test set is required before concluding that performance is stable across the broader Serdos domain. Cross-collection neural-IR studies and reproducibility research provide important guidance for such extensions (Marchesin et al., 2020; Maistro et al., 2023). While heterogeneous retrieval benchmarks caution against assuming that one retriever configuration transfers uniformly across domains (Thakur et al., 2021). From a technical perspective, future development could also incorporate visual information such as slide OCR so that material appearing only on screen and not spoken aloud is not excluded from the index.

Conclusion

This study implements an NLP-based hybrid retrieval pipeline for locating timestamped segments in one Indonesian-language Lecturer Certification video. Faster-Whisper medium produced 748 ASR segments that were then formed into 60 chunks using a 40-second sliding window with a 10-second overlap. ASR evaluation yielded WER of 14.76% and CER of

7.23%. At α 0.50, Precision@1 of 0.6000 shows that 12 of 20 queries placed a relevant chunk at rank one, while Recall@10, MRR@10, and nDCG@10 were 0.9208, 0.7288, and 0.7150, respectively. Among the five tested configurations, α 0.50 produced the highest Precision@1 and MRR@10, while α 0.75 produced the highest nDCG@10 of 0.7273; therefore, no single α was superior across all metrics. These findings demonstrate the feasibility of the pipeline as a video-segment navigation mechanism within the evaluated scenario, but they do not establish universal superiority of hybrid retrieval. Limitations include one video, 20 queries, ground truth validated by one researcher, binary relevance, a single chunk configuration, reuse of the same queries for tuning and evaluation, and the absence of visual-content analysis. Future work should expand the dataset, involve independent annotators, separate validation and test sets, compare chunking configurations, and evaluate reranking and visual information.

Acknowledgment

-

References

Ahlawat, H., Aggarwal, N., & Gupta, D. (2025). Automatic speech recognition: A survey of deep learning techniques and approaches. *International Journal of Cognitive Computing in Engineering, 6*, 201–237. https://doi.org/10.1016/j.ijcce.2024.12.007

Alabbad, M., & Alhoshan, W. (2026). Automatic speech recognition in healthcare in the post-LLM era: A scoping review. *Healthcare, 14*(10), 1333. https://doi.org/10.3390/healthcare14101333

Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-text semantic similarity (STSS): Techniques, challenges and future perspectives. *Applied Sciences, 13*(6), 3911. https://doi.org/10.3390/app13063911

Arazzi, M., Ferretti, M., & Nocera, A. (2023). Semantic hierarchical indexing for online video lessons using natural language processing. *Big Data and Cognitive Computing, 7*(2), 107. https://doi.org/10.3390/bdcc7020107

Bruch, S., Gai, S., & Ingber, A. (2024). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems, 42*(1), Article 20, 1–35. https://doi.org/10.1145/3596512

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2026). The Faiss library. *IEEE Transactions on Big Data, 12*(2), 346–361. https://doi.org/10.1109/TBDATA.2025.3618474

Fan, Y., Xie, X., Cai, Y., Chen, J., Ma, X., Li, X., Zhang, R., & Guo, J. (2022). Pre-training methods in information retrieval. *Foundations and Trends in Information Retrieval, 16*(3), 178–317. https://doi.org/10.1561/1500000100

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., Croft, W. B., & Cheng, X. (2020). A deep look into neural ranking models for information retrieval. *Information Processing & Management, 57*(6), 102067. https://doi.org/10.1016/j.ipm.2019.102067

Gupta, D., Attal, K., & Demner-Fushman, D. (2023). A dataset for medical instructional video classification and question answering. *Scientific Data, 10*, 158. https://doi.org/10.1038/s41597-023-02036-y

Han, M., Zhang, X., Yuan, X., Jiang, J., Yun, W., & Gao, C. (2021). A survey on the techniques, applications, and performance of short text semantic similarity. *Concurrency and Computation: Practice and Experience, 33*(5), e5971. https://doi.org/10.1002/cpe.5971

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems, 20*(4), 422–446. https://doi.org/10.1145/582415.582418

Kiss, C., Nagy, M., & Szilágyi, P. (2025). Max–Min semantic chunking of documents for RAG application. *Discover Computing, 28*, Article 117. https://doi.org/10.1007/s10791-025-09638-7

Liu, M., Nie, L., Wang, Y., Wang, M., & Rui, Y. (2023). A survey on video moment localization. *ACM Computing Surveys, 55*(9), Article 188, 1–37. https://doi.org/10.1145/3556537

Maistro, M., Breuer, T., Schaer, P., & Ferro, N. (2023). An in-depth investigation on the behavior of measures to quantify reproducibility. *Information Processing & Management, 60*(3), 103332. https://doi.org/10.1016/j.ipm.2023.103332

Marchesin, S., Purpura, A., & Silvello, G. (2020). Focal elements of neural information retrieval models: An outlook through a reproducibility study. *Information Processing & Management, 57*(6), 102109. https://doi.org/10.1016/j.ipm.2019.102109

Masuyama, M., Kawahara, T., & Matsuda, K. (2024). Video retrieval system using automatic speech recognition for the Japanese Diet. In *Proceedings of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN) @ LREC-COLING 2024* (pp. 145–148). ELRA and ICCL. https://aclanthology.org/2024.parlaclarin-1.21/

Moffat, A. (2022). Batch evaluation metrics in information retrieval: Measures, scales, and meaning. *IEEE Access, 10*, 105564–105577. https://doi.org/10.1109/ACCESS.2022.3211668

Novriansyah, F., Prasti, D., & Uramako, E. (2023). Pembuatan Video Profil Pemodelan Tata Ruang 3 Dimensi Gedung SMAN 15 Luwu . *Journal Artificial: Informatika Dan Sistem Informasi*, *1*(1), 53–59. https://doi.org/10.54065/artificial.250

Polat, H., Turan, A. K., Koçak, C., & Ulaş, H. B. (2024). Implementation of a Whisper architecture-based Turkish automatic speech recognition (ASR) system and evaluation of the effect of fine-tuning with a low-rank adaptation (LoRA) adapter on its performance. *Electronics, 13*(21), 4227. https://doi.org/10.3390/electronics13214227

Putra, S. J., Wijaya, A., & Alam, R. G. G. (2025). Analysis of MP3 bitrate on the accuracy of academic audio transcription using Whisper large-v3. *Jurnal Sistem Cerdas, 8*(2), 160–168. https://doi.org/10.37396/jsc.v8i2.528

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202, pp. 28492–28518). PMLR.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Reimers, N., & Gurevych, I. (2020). Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 4512–4525). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.365

Riskawati, R., & Saleh, M. (2026). Implementasi Project-Based Learning Berbantuan Tinkercad untuk Mendukung Pemahaman Konsep Listrik Dinamis Mahasiswa. *Journal Artificial: Informatika Dan Sistem Informasi*, *4*(2), 227–238. https://doi.org/10.54065/artificial.1547

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management, 24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., & Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*.

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., & Bhuiyan, A. (2024). Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys, 56*(7), Article 185, 1–33. https://doi.org/10.1145/3648471

Wu, J., Liu, W., Liu, Y., Liu, M., Nie, L., Lin, Z., & Chen, C. W. (2026). A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 48*(2), 1521–1541. https://doi.org/10.1109/TPAMI.2025.3615586